Forecast the number of burglaries in 44th district of the NYC

By Kateryna Brodiansky

1. In the first part of this exercise, I studied the data and visualized them using graphs and maps. As you can see in the attached notebook: "NYC_crimes_Part2.". I analyzed the types of variables, the distribution of data in space and time, the number of burglaries, depending on time such as year, month and day of the week, and depending on the area. Which population groups commit crimes of this type and more…

I checked the number of missing values and implemented actions to replace or delete them based on each specific variable.

There are no outliers in this dataset where we need to check, most of the data is categorical. We check the location data with graphical visualization, and we limit the data by time to the range that we need. The results in more detail I comment on the notepad. I used home burglaries data subset of all New York City and subset Precinct no. 44 and kept comparing them all the time.

The results were very interesting and informative and helped me choose a solution strategy.

2. For the forecast, I grouped already cleaned data on the date of the crime and the district number. Because the aim of this forecast is the number of burglaries in a particular district on specific days.

I tested four machine learning models: GradientBoosting, KNeighbors, and RandomForest for forecasting and validated using the mean square error metric.

3. I chose a GradientBoosting model with minimal MSE = 4.45831e-08
and made a forecast for all the districts for each day. They give me better results than the forecast by one district.

4. I attach a table with the number of home burglaries at precinct no. 44 during the first 3 months of 2018 by days. In this area during this period was committed 47 home burglaries.

5. As a result of research, I managed to get the result with high accuracy, so I could advise using this methodology to solve this specific problem. But I can tell what I planned to do if the result did not meet the required accuracy. I would apply cross-validation with a period of one year and be validated in the first three months of each year. The next step would be to change the grouping

periods and or CV periods. I also tried to enrich the data using other information like a population and education by district, maybe holidays. I would also try to the tuning of parameters in the selected model…..

6. As you can see from the graphs in the first part of my technical solution, it is a neighborhood of the city that has the strongest influence on the number of home burglaries. If it's really important for you, I would recommend you choose an apartment in the Staten Island area. In the process of this work, I would check the following relations:

- analysis of the number of burglaries by neighborhood
- relation by the count of population and the number of all crimes of and specifically of this type
- analysis of other characteristics of the neighborhood as the level of education and unemployment and checking the relationship with the number of home burglaries.
- quantity forecast of home burglaries for each district
- analysis of the probability for each person that in his apartment there will be a home burglary depending on the neighborhood.